VARBRULプログラムとは何か*

松田謙次郎

VARBRUL Primer

MATSUDA Kenjiro

Abstract

VARBRUL is a logistic regression program that has been used in variationist sociolinguistics since the 1970s. Although it is by now the default analysis tool in the field, the software is far from popular in Japanese sociolinguistics,partly due to the poverty of introductory texts for this sophisticated freeware. This paper is intended to be a general introduction to the software, covering the statistical problems it is intended to deal with, as well as the practical how-to's of the program.

VARBRULプログラムは欧米の変異理論研究では一般的なフリーウェアだが、日本においてはVARBRUL登場後30年たってなお、その紹介が十分に成されているとは言い難い状況にある。本稿ではVARBRULの基本から、実際の使用法、そして現バージョンのVARBRULが持つ諸問題点を指摘するものである。

1. はじめに

「VARBRULプログラム」とは、Cedergren and Sankoff (1974)、またその前年のCedergren (1973)において初めて使用された言語変異分析用フリーウェアの名称である。日本ではあまり知られているとは言い難いが、欧米の変異理論研究論文では頻繁に登場し、論文の議論もその出力結果に基づいてなされることが多い。しかしながらそうした状況にある欧米においても、長い間このプログラムに関する本格的な教科書は存在せず、ソフト付属のマニュアルや概説書、また変異理論最大の学会であるNWAV(New Ways of Analyzing

*本稿は、2004年9月5日に東京大学本郷キャンパスで開催された、第14回社会言語科学会大会ワークショップ「ことばのバリエーションをつかまえる: 分析ツールとしてのVARBRULプログラムの活用法と隣接分野への応用」で口頭発表したものに大幅に加筆・修正を加えたものである。ワークショップの企画者である太田一郎、企画段階からコメントを頂いた朝日祥之、片岡邦好、高野照司の諸氏、および当日コメントを頂いた参加者諸氏に謝意を表したい。

Theoretical and Applied Linguistics at Kobe Shoin 9, 31–56, 2006.
© *Kobe Shoin Institute for Linguistic Sciences.*

Variation) を始めとする学会で付随して開催されるワークショップなどでその使用法を学んでいたのが実情であった。Paolillo (2002) の出版によって、変異理論界は、ようやく最近統計的背景や歴史までをも含めた、VARBRUL プログラムの全貌を纏めた教科書を持つことになったと言っても過言ではない。

実際に現在使用されているソフトは、VARBRUL (MS-DOS)、Goldvarb 2.1 (Macintosh)、GoldVarb2001 (Windows) および Goldvarb X （Windows および Macintosh）とあるが、変異理論関連の論文では、一概に VARBRUL analysis と言うことが多い。[1] 現在、VARBRUL プログラムは以下のウェブサイトからダウンロードが可能である。

- アメリカ・ペンシルバニア大学言語データコンソーシアム (VARBRUL) `ftp://ftp.ldc.upenn.edu/pub/ldc/misc_sw/varbrul.tar.Z`
- カナダ・ケベック大学モントリオール校数学研究センター (Goldvarb 2.1) `http://www.crm.umontreal.ca/~sankoff/GoldVarb_Eng.html`
- イギリス・ヨーク大学言語学科 (Goldvarb2001) `http://www.york.ac.uk/depts/lang/webstuff/goldvarb/`
- カナダ・トロント大学タリアモンテ教授ウェブサイト (Goldvarb X) `http://individual.utoronto.ca/tagliamonte/goldvarb.htm`

もっとも古い DOS バージョン以外のプログラムは、細かな違いはあるがほぼ同機能であり、もちろん GUI インターフェースを持つ。この論文では Goldvarb X に沿って解説を進めるが、基本的な部分では全く変わらないと思って良い。

2. 東京語「を」ゼロマーク化データ

まず最初に、表 1–2 に纏められるようなデータがあるとしよう。このデータは、Matsuda (1995) および松田 (2000) による東京方言の対格格助詞「を」のゼロマーク化に関するもので、東京方言話者 39 名の自然談話より得られた 7,529 件のデータである。。データは、「を」がゼロマーク化されたか否か（つまり「を」として実現されたか）をコードした従属変数と、助詞の付く目的語名詞の形式（語彙名詞、節、WH 代名詞、代名詞）、およびその名詞と動詞が隣接しているかどうかをチェックした隣接性の 2 つの独立変数からなる。表 1 にコード表を掲げておく。[2]

例えば以下のような発話であれば、最初の発話は `11SA` と、2 番目の発話は `02FN` とコードされることになる:

(1) でさー、あいつがついに携帯買ったんだって！

[1] 以後、この論文で「VARBRUL プログラム」と言う場合、それは「バージョンを問わずすべての VARBRUL プログラムで」という意味である。

[2] 実際には他の多くの要因がコードされているが、ここではすべて割愛する。詳しくは Matsuda (1995), 松田 (2000) を参照のこと。

(2) それで、あの先生はあの店に出ていた携帯を急いで買われました。

さて、今回の全データ 7,529 件を表にまとめると、表 2 のようになる。[3]

このデータから、各独立要因が「を」のゼロマーク化に対して持つ貢献度（重み）を割り出し、どの要因がどれほど影響を与えているのかを見極めたいとしよう。実はこうした状況は、何らかの形で言語変異や言語変化の分析をしているとしばしば遭遇するものである。現象や各要因を適当に変更してみれば、こうした場面が以下に一般的なものかがわかるだろう。

2.1 表の目視によるパーセンテージの分析

まず浮かぶのは、各セルに記入されているパーセンテージを目視により比較して、各要因の相対的な重みを割り出そうとする考え方である。これはもっとも原始的な方法だが、要因数が 2（すなわち従属変数一つと独立変数一つ）～3 位であれば十分可能である。要因が増えて全体像が把握しにくい場合は、要因のレベルによって別表を作成すればよい（これを「層別」と言う）。表 2 でもスタイルにより表を大きく 2 つに分割している。この表からは、少なくとも次のようなことが容易に読み取れるだろう:

(3) パーセンテージから読みとれること

- 隣接環境の方が、非隣接環境よりもゼロマーク化が起きやすい。
- 名詞形式差では、ゼロマーク化されやすい順に WH 代名詞 > 語彙名詞 > 代名詞 > 節 という序列があり、これはスタイルや隣接性のいかんに関わらず同様である。
- スタイル差では、ゼロマーク化は改まったスタイルよりもくだけたスタイルで起きやすく、この関係も隣接性のいかんに関わらず成立する。

ただし、いずれにしてもこの方法ではあくまでセル間のパーセンテージの比較でしかないので、要因の持つ統計的有意性はわかりようがない。また、変数が増加した場合、すぐに手に負えなくなる。このデータの場合、独立変数同士は独立の関係にあるので、表からの読み取りも比較的簡単だが、いつもこうだというわけではない。特に、性別・年齢など話者の社会的属性が組み込まれた場合、社会的要因同士の関係は独立でないことが多い。[4] 最後の問題点として、層別化した場合、後述する「シンプソンのパラドックス」に直面する事態が考えられることも指摘しておこう。

2.2 グラフによる分析

パーセンテージの目視による比較をもう少し工夫すると、グラフによって視覚化して比較を行うという発想に至る。2～3 要因を組み合わせるなどしてグラフを作成すれば、データのおよそのあらましは把握できる場合が多い。そもそもグラフ作成は、どのような

[3] 0 は格助詞「を」がゼロとして実現されること（ゼロマーク化）を示す。

[4] 興味深いことに、言語変異においては、いわゆる言語内的要因同士はたいてい独立であること、言語内的要因と外的要因（=社会的属性）もたいてい独立であること、そして外的要因同士は交互作用が見られることも多いことがわかっている。

統計分析を施すことになろうとも、まずはデータ分析の第一歩となるべきであり、データ中のパターンを発見するという探索的見地からも、グラフによる分析は決して間違いではない。

このデータの場合も、データをグラフ化すると図1のようになり、ここからもある程度の読み取りは可能である。しかしながら、問題はパーセンテージの場合と同様である、要因が4,5,…と増加するに従い、グラフによる分析は困難なものになり、各要因の効き具合を細かく吟味するのは不可能とまでは行かなくとも、きわめて骨の折れる作業となろう。また、それにしてもある要因のレベル間の区別や要因そのものの有意性は、グラフのみからでは調べようもない。

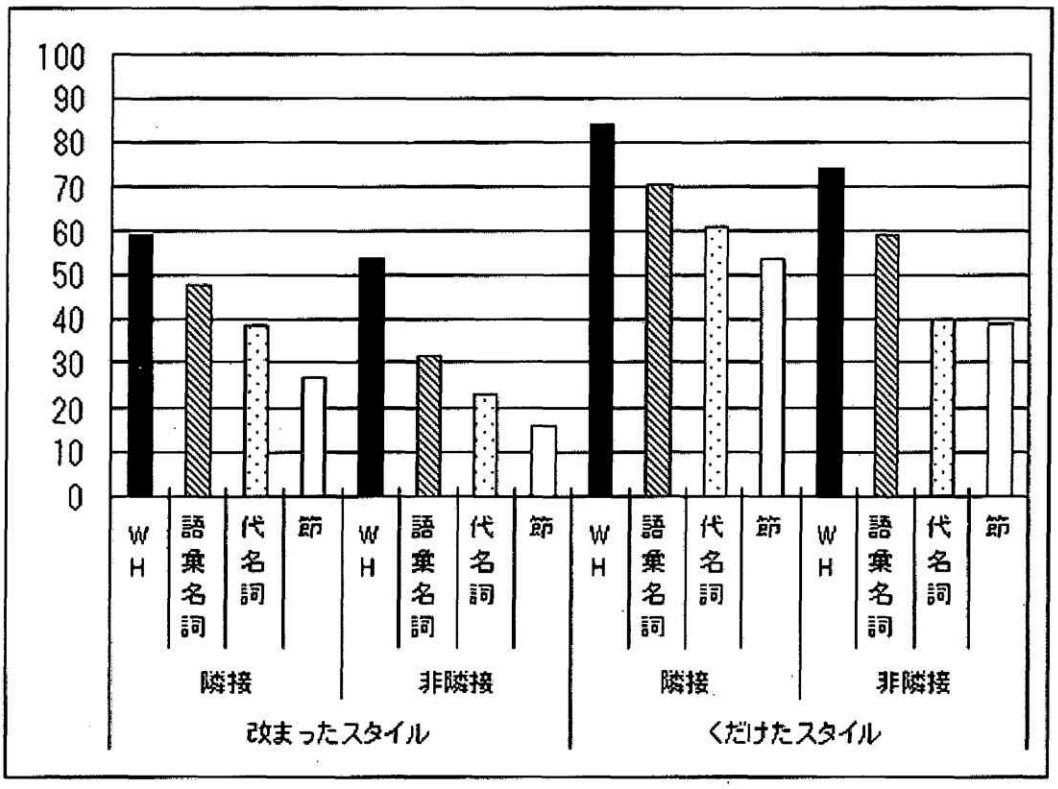


図1: 東京方言データ（表2）のグラフ化

2.3 統計的検定: カイ自乗検定

パーセンテージのような記述統計的発想からさらに発展すると、次は検定を含んだ統計的解析ということになる。データは名義尺度データなので、カイ自乗検定がすぐに浮かぶ。目的変数であるゼロマーク化の有無の頻度と各要因を一つづつ掛け合わせた分割表でカイ自乗値を計算し、自由度を参照しながらその有意性を検討するという手順を踏む。有意であった場合、その要因はゼロマーク化の実現（つまりはゼロ）に何らかの形で関連しているという結論を導き出すことができる。例えばこのデータで3つの独立変数

それぞれにカイ自乗検定を試みると、[5] すべてが $p < 0.001$ レベルで有意である (表 3–5)。

こうしたカイ自乗法を用いた分割表分析は、言語学の論文でも目にすることが多い。しかしながら、実はカイ自乗を用いた分析には問題点がいくつかある。まず、通常のカイ自乗法ではその分割表全体の分布については検定が可能だが、各要因のレベル間については何も言うことがない、という点である。つまり、パーセンテージから大体どこら辺に大きな差がありそうかはわかるが、はっきりと「レベル 1 とレベル 2 の区別は有意だが、レベル 2 とレベル 3 の区別は有意ではない」といった情報は与えてくれないのである。これだと、せっかくその要因が重要だと言うことがわかりながら、そこから先の細かなより分析に進めることができないことになってしまう。

実は、カイ自乗を発展させた形として残差分析を行ない、有意差検定を行うことでクロス表に見られる分布特徴をより確かな形で把握することは可能である (Everitt, 1992)。残差分析では、各セルの観測値と予測値の差を予測値の平方根で割り標準化残差を計算し、さらにこれを標準偏差で割ることで得られる調整化残差（adjusted standardized residual, ASR）を使用する。この値は、平均 0、標準偏差 1 で正規分布に近似するため、その値の絶対値が 1.96 を超えれば 5%域で有意な相関を持つ組み合わせ（＝セル）を見つけることができる (Haberman, 1973)。ここから表全体のカイ自乗検定の有意性に、どのセルが貢献したのかを点検し、そのパターンを見ることで表の特徴を捉えるわけである。[6]

しかし、残念ながら残差分析によっても解決しない、単変量解析特有の問題が存在する。それは交絡要因 (confounding factor) の問題である。交絡要因とは、ある現象 X に関わる真の要因 A が、別の要因 B（交絡要因）の存在によりその関わりが見えにくくなり、表面的に要因 B が現象 X の重要な要因とみなされてしまうようなケースである。たとえば発癌要因とされるものには、飲酒、喫煙、肥満、などといった因子が知られているが、一定地域の住民に調査をした場合、こうした要因で構成されるクロス表すべてのセルに均等にデータが分布するとは考えられない。そこにはたいてい何らかの偏りが出るはずだが、こうした場合に癌の有無に対して、飲酒のみあるいは喫煙のみでまとめた表を作り、その効果を論じる場合を考えてみれば、その危険性は明らかであろう。本当のところはいずれの要因がどれほど効いていたのかを正当に判断するには、飲酒、喫煙、そしてその他も含めたすべての要因を一度に評価し、他の要因をコントロールした上である要因の効果を計算する必要があるわけである。自然発話に基づくデータを扱う変異理論でも事情は同じであり、性別、年齢、学歴といった社会的要因は言うまでもなく、今回のデータのような文法的カテゴリーについても分布は一様ではない。こうした分野ではあらかじめ実験計画を立て、要因をコントロールした上でサンプル収集を行う実験データとは異なる種類の困難がつきまとうわけである。

この交絡要因の極端なケースとして、「シンプソンのパラドックス（Simpson'sparadox）」として知られる現象を紹介しておきたい (Simpson, 1951)。これは、一口に言うと層別ク

[5] スタイルと隣接性については 2×2 表になるため、イェーツの連続性のための修正法に従って計算してある。ただし、これによる差は各々の場合で 1 以下であり、結果にはほとんど影響を与えていない。

[6] なお、Excel でこれを計算する場合、内田 (2000) が参考になる。

表 1: 東京語「を」ゼロマーク化: コード表

要因	レベル	コード
名詞句形式差	語彙名詞	1
	節	2
	WH	3
	代名詞	8
スタイル	改まったスタイル	F
	くだけたスタイル	S
隣接性	隣接	A
	非隣接	N

表 2: 東京語ゼロマーク化データ （上段: 改まったスタイル; 下段: くだけたスタイル）

	改まったスタイル・隣接				計	改まったスタイル・非隣接				計
	語彙	節	WH	代名詞		語彙	節	WH	代名詞	
∅	840	45	30	160	1,075	132	13	8	41	194
総数	1,759	168	51	417	2,395	414	81	15	178	688
%∅	48%	27%	59%	38%	45%	32%	16%	53%	23%	28%

	くだけたスタイル・隣接				計	くだけたスタイル・非隣接				計
	語彙	節	WH	代名詞		語彙	節	WH	代名詞	
∅	1,808	82	63	357	2,310	420	32	20	95	568
総数	2,573	153	75	587	3,388	713	82	27	236	1,058
%∅	70%	54%	84%	61%	68%	59%	39%	74%	40%	54%

表 3: 名詞句形式差によるゼロマーク化の分布 ($X^2(3) = 168.551, p < 0.001$)

	語彙名詞	節	WH	代名詞	計
を	2,259	312	47	765	3,383
ϕ	3,200	172	121	653	4,146
計	5,459	484	168	1,418	7,529

表 4: スタイルによるゼロマーク化の分布 ($X^2(1) = 407.068, p < 0.001$)

	改まったスタイル	くだけたスタイル	計
を	1,814	1,569	3,383
ϕ	1,269	2,877	4,146
計	3,083	4,446	7,529

表 5: 隣接性によるゼロマーク化の分布 ($X^2(1) = 120.509$, $p < 0.001$)

	隣接	非隣接	計
を	2,398	985	3,383
ϕ	3,385	761	4,146
計	5,783	1,746	7,529

ロス表において、各層別のクロス表で成立する関係と、層を解消してすべてを合計したクロス表で成立する関係が異なってしまう事態のことを言う。Fienberg (1980, 50ff.) による例を見てみよう（表 6–7、原著を一部改変）。

表 6: シンプソンのパラドックス (1): それぞれの層の表

	C		
	B'	B	計
A'	950 (95%)	9,000 (90%)	9,950
A	50 (5%)	1,000 (10%)	1,050
計	1,000 (100%)	10,000 (100%)	11,000

	C'		
	B'	B	計
A'	5,000 (50%)	5 (5%)	5,005
A	5,000 (50%)	95 (95%)	5,095
計	10,000 (100%)	100 (100%)	10,100

ここでは A, A' という 2 レベルからなる第 1 要因、同様に B, B' からなる第 2、C, C' を持つ第 3 要因があるとして、第 3 要因で層別化した 2 つの表（表 6）と、これらを合併した表（表 7）がある。層別化されたそれぞれの表では、いずれも [BA のセル] > [$B'A$ のセル] という関係が成立しているが、合併した表ではこれが逆転して、[BA のセル] < [$B'A$ のセル] となってしまっている。

シンプソンのパラドックスが教えてくれるのは、やはり複数要因が関わる現象において、単変量分析の繰り返しで対応していると、見落としてしまうパターンがあるということであり、それを避けるためには、複数要因が関わっているならそれらの要因を一度に見なさいということである。そこで、多重クロス表を一度に検討する手続きを統計的にやってくれる多変量解析が必要になるのである。[7]

3. ロジスティック回帰分析とは

多変量解析と一口に言っても、分析の目的や独立・従属変数の尺度などによってさまざまな手法がある。今回の場合、やりたい分析は独立変数が名義尺度、連続尺度の場合

[7]実はカイ自乗には独自の問題もある。それは、サンプル数が大きくなりすぎるとほとんどの場合に有意な差が出てしまい、本質的でないカテゴリの変化が、分布のちがいを示す原因を作ってしまうという問題である (西平, 1985, 158)。

表 7: シンプソンのパラドックス (2): 要因 C について合併された表

	B'	B	計
A'	5,950 (54%)	9,005 (89%)	14,955
A	5,050 (46%)	1,095 (11%)	6,145
計	11,000 (100%)	10,100 (100%)	21,100

の回帰分析ということになる。実際、多くの言語変異・変化現象の場合は従属変数（上の例で言えばゼロマーク化がが起きる/起きない）が名義尺度であることが多い。[8] 従属変数が名義尺度の場合、単純最小自乗法（Oridinary Least Squarex Method, OSL）や重み付き最小自乗法（Weighted Least Square Method, WLS）に基づく通常の回帰分析では都合が悪い。それは、こうした場合、予測対象が 0 から 1 の間に収まる割合（生起確率）で表現されることになるわけだが、回帰モデルは時にこの範囲を逸脱する数字を予測としてはじき出してしまうからである。

ロジスティック回帰分析 (logistic regression analysis) は、こうした独立変数・目的変数共に名義変数の場合の回帰分析なのである。この手法では、生起確率そのものを予測するのではなく、生起確率 p のロジット $(ln(\frac{1-p}{p}))$ が予測対象となる。ロジットとは、現象の生起確率 p のオッズ $(\frac{1-p}{p})$ を自然対数変換したものである。ロジット化された、多重クロス表のそれぞれのセルの確率の予測をする中で、各々の要因の重みを算出して、最終的に予測モデルを作ろうとするわけであり。そして、それをやってくれるソフトの一つが VARBRUL なのである。[9]

ロジスティック回帰は上述の事情もあって、当初疫学分野で大きく発展した統計手法である。[10] アメリカの心臓疾患に関する著名な大規模調査であるフラミンガム心臓病調査プロジェクト (Framingham Heart Study) では、血中コレステロール値、血圧、喫煙習慣の有無などを始めとするさまざまな要因から心臓疾患に関わる要因の同定を試みたが、この中で心臓疾患の有無（1/0）を複数の名義（例: 喫煙習慣の有無）・連続変数（例: 血中コレステロール値、血圧など）から予測する統計手法が必要になり、ロジスティック回帰の開発につながった。同時に進行した大型計算機の一般化と共にこの手法は広く経

[8] もちろんこれがすべての場合だというわけではない。たとえば母音の変異・変化を扱おうとすれば、F1, F2 などのフォルマント周波数が従属変数になるが、これなら通常の重回帰分析が可能である。こうした研究については、ラバブの一連の母音推移に関する研究が参考になる (Labov, 1994)。なお、ロジスティック回帰では、独立変数についてはカテゴリー変数と共に連続変数を含むことも可能であるが、こうしたモデルでのパラメータ推定を Excel で行う方法が増山・山田 (2004) にあり、参考になる。

[9] VARBRUL が行うのはロジスティック回帰およびクロス表作成だが、ロジスティック回帰に近い内容を持つ解析法として対数線形分析 (log-linear analysis) が知られている。しかしながら両者は明確に異なるものである。対数線形分析は簡単に言えば従属変数と独立変数を持った回帰ではなく、名義変数からなるクロス表のセル度数を予測するモデルだと思えばよい。「名義変数からなるクロス表」なので、対数線形分析では、たとえば連続変数をそのままモデルに組み込むことはできない。一方、ロジスティック回帰ではそれが可能である。VARBRUL でそれができないのは、単に VARBRUL プログラムの機能が限定されているからに過ぎない。VARBRUL プログラムを対数線形分析ソフトと誤解してのではと思われる解説も見かけるので（たとえば前川 (2001)）、注意が必要である。

[10] ロジスティック回帰に深い関連を持つ対数線型モデルも、アメリカの麻酔薬研究 (National Halthane Study) からの発展である。

済学、社会学など医学以外の分野にも応用されるようになったという経緯がある。それが、70 年代になりやっと言語学にまで及んできたわけである。手法自体は汎用統計パッケージとして知られる SPSS や SAS にも含まれており、現在ではもちろんパソコン上で手軽に分析可能である。変異理論的分析用にでこれを手軽に扱えるパッケージにしたのが、VARBRUL プログラムなのである。[11]

3.1 VARBRUL 分析による方法

さて、それではロジスティック回帰ソフトである VARBRUL では、このデータに対してどのような分析をしてくれるだろうか。細かいことを追う前に前に、まずはゼロマーク化データの VARBRUL プログラムの最新版である Goldvarb X による分析結果を見てみよう（表 8）。

表 8: ゼロマーク化データの Goldvarb X による分析結果

要因	レベル	重み	p 値
名詞句形式差	語彙名詞	0.532	$p < 0.001$
	節	0.334	
	WH	0.687	
	代名詞	0.415	
スタイル	改まったスタイル	0.359	$p < 0.001$
	くだけたスタイル	0.599	
隣接性	隣接	0.535	$p < 0.001$
	非隣接	0.385	
定数項		0.554	

重み（パラメータ値）は、0 から 1 の間の値を取り、0.5 を境に 0.5 より大きいとその要因/レベルがゼロマーク化を実現する方向に効果があることを示し、逆に 0.5 以下だとゼロマーク化を起こさない方向に働くものと解釈される。このことを踏まえると、先にパーセンテージから読みとったことがすべて正しかったことがまず読みとれる。また p 値を見て行くと、やはり先のカイ自乗検定の結果を裏付けていることもわかる。ならばパーセンテージやカイ自乗で十分ではないか、という批判が出るかもしれないが、上の結果がそれらと決定的に違うのは、これが「他要因の効果を考慮した結果」だという点なのである。すでに見た通り、パーセンテージのデータからだけでは、各要因に独立してどれほどの効果があるのかは分かりようもなく、要因一つ一つののカイ自乗検定からでは、果たして他要因を考慮してもその要因が有意なのかも判別できないのである。ロ

[11] 日本語で書かれたロジスティック回帰分析自体の解説書は、高橋 (1995), 丹後, 山岡, 高木 (1996), 浜島 (2000) など医療統計学関係が目立つ。言語学関連では、先に挙げた Paolillo (2002) 以外に Rietveld and van Hout (1993) があり、政治学の増山・山田 (2004) も扱いは短いがわかりやすい。英語であれば、Fienberg (1980) や Everitt (1992) が参考になる。後者には初版に邦訳がある (エヴェリット, 1980)。SPSS でロジスティック回帰を走らせる場合の簡単な解説であれば、石村 (2001) が手軽である。

ジスティック回帰の結果は、たとえばスタイル差や名詞句の形式差の効果を勘定に入れても、隣接性要因は高度に有意な効果をゼロマーク化に対して持つことを示しているのである。

さらに名詞句形式差については、例えば代名詞とWH代名詞は「代名詞」として同じカテゴリーにまとめるのは適当でないことや、語彙名詞と代名詞を一つのカテゴリーにまとめるのは、いずれも統計的見地からは適当でないことも、追加分析から確認可能である。これについても他要因をコントロールした上でのことであり、カイ自乗の残差分析から分かることよりもはるかに多くのことを教えてくれていることは言うまでもない。

こうして得られた予測モデルはどれくらい正確な予測をしてくれるだろうか。表8のパラメータを使ってゼロマーク化の生起度数を予測した結果が表9である。一見して明らかなとおり、きわめて観測値に近い予測値をはじき出している。

表9: 観測値とGoldvarbによる予測値（上段: 改まったスタイル; 下段: くだけたスタイル）

	改まったスタイル・隣接				改まったスタイル・非隣接			
	語彙	節	WH	代名詞	語彙	節	WH	代名詞
∅	840	45	30	160	132	13	8	41
予測値	837.629	48.180	32.488	151.148	136.865	14.521	7.321	42.003

	くだけたスタイル・隣接				くだけたスタイル・非隣接			
	語彙	節	WH	代名詞	語彙	節	WH	代名詞
∅	1,808	82	63	357	420	32	20	95
予測値	1820.605	79.097	61.775	353.440	404.939	30.147	19.367	106.476

4. VARBRUL分析の流れ

ここで、VARBRUL分析の全体的な流れを見てみよう。VARBRUL分析は、トークンファイル（コーディング）の作成が最初である。次に要因指定をするコンディションファイルを作成し、後々の分析の元となる中間的データファイルであるセルファイルの作成がこれに続く。セルファイルは、次のステップであるクロス表とロジスティック回帰への入力ファイルとなる。分析がもっとも順調に進行した場合、直線的にロジスティック回帰に進んで結果が出てゴールとなるが、ほとんどの場合こうはならない。実際の分析では、出てきたロジスティック解析やクロス表の結果を睨みながら、再びコンディションファイルに戻ってセルファイルを作り直し、もう一度クロス表やロジスティック回帰に向かい、満足な結果が出るまでこのサイクルを繰り返すという、循環的作業を繰り返すことになる。場合によっては、最初のトークンファイルも分割する必要が出てくるなど、分析過程で無数のファイルが生じるため、あとでどのファイルにどのような分析を考えていたかがたどれるよう、入念なログ取りが欠かせない。次セクションから、各ステップを細かく見てみよう。

4.1 コーディング: トークンファイル

あらゆる統計ソフトの最初のステップは入力データの作成であるが、Goldvarb もその例外ではない。Goldvarb の最も基本的なデータファイルはトークンファイルと呼ばれるファイルであり、(4) のような内容である。

(4) トークンファイルの例

```
(18FA
(11FA
(18FA
(18FA
(11FA
(11FA
```

トークンファイル作成には Goldvarb 付属の機能を使っても良いが、テキストフォーマットのデータであれば読み込めるので、テキストエディタで作成しても良いし、Excel で作成したものをテキストフォーマットにして読み込んでも良い。トークンファイルのフォーマットには、以下のようなルールがある。

(5) トークンファイルのフォーマット

- 文字はすべて半角で入力。
- 一件一行に入力。
- 各行は開き括弧で始める。
- 第 2 コラムは従属変数で、以降独立変数を入力する。
- 各行のコード連鎖の後ろに、スペースを空けてテキスト入力が可能。
- セミコロンでコメントが入力可能

各行が開き括弧で始まる規則については、Excel で括弧のコラムを一気に記入するなり、エディタのマクロ機能を使うなりして能率化する方がよいだろう。

こうして入力したデータに加えて、次にデータチェックのために要因詳細指定作業 (Factor Specification) を行う。これは、要因[12] ごとにレベルとデフォルトの値を入力し、データがきちんとこの通りに入力されているのかをチェックするのである。これが終了すると、ファイル末尾に要因ごとにデフォルト値とレベルが記入されてトークンファイルが完成する (図 2)。

なお、すでに Excel で分析を進めていたファイルを整形しテキストファイルとして読み込む場合や、すでに事前の分析から要因詳細指定作業が不要と判断できるのであれば、メニューの Tokens から Generate Factor Specification へと進み、自動的に現在のファイルから要因詳細指定を割り出し、そのままトークンファイル作成終了へと進むことも可能である。

[12] VARBRUL 関連の資料全体に言えることだが、一般に統計学では要因 (factor) は独立変数と同義に使われ、各要因の中に下位区分としてレベルが存在するという言い方をする。これに対して、VARBRUL 関連の著作では、要因を *factor group* と呼び、レベルを *factor* と呼ぶのが通例で、注意が必要である。

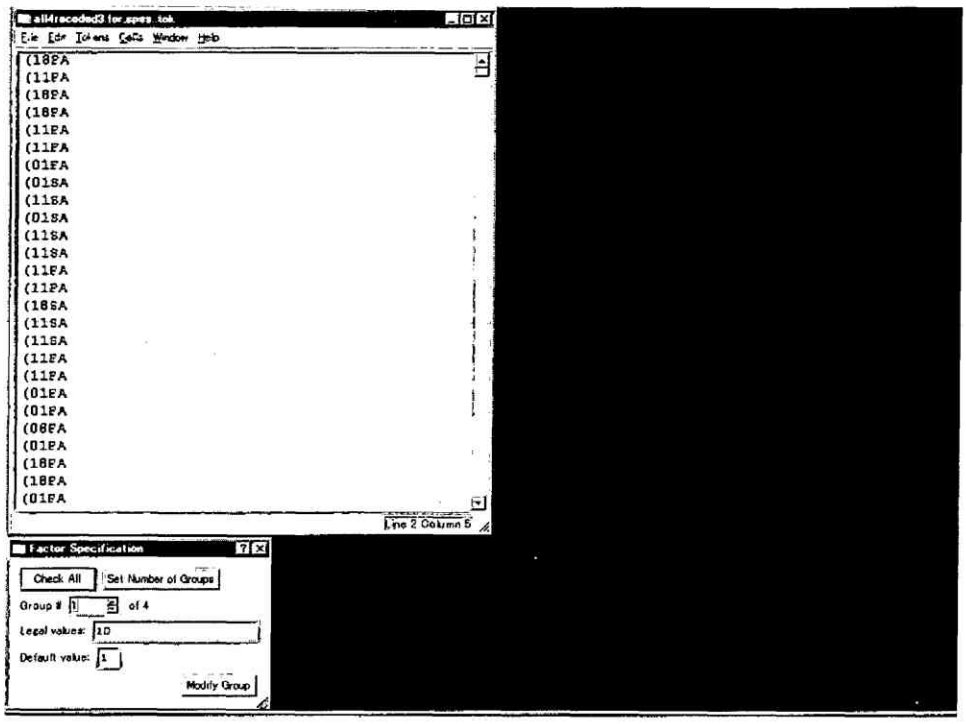


図 2: トークンファイル

4.2 コンディションファイル

トークンファイルに加えて、VARBRUL を動かすにはもう一つの入力ファイルが必要である。それがコンディションファイルであり、トークンファイルのデータを、分析に使いたい変数だけ取り出したり、変数のレベルをさまざまに結合させるなど、必要に応じてまとめ直すための指示をここで記録する。コンディションファイルもテキストファイルであり、拡張子に .cnd を持つ。ここで記録されなかった要因については、これ以降の分析は全くなされない。この記述方法もトークンファイルのそれと似通っており、セミコロンの後にコメント可能であり、ファイルの最初のコラムは "(" で始まり、ファイルの最後は ")" である必要がある。[13] 分析の初期段階では、各要因でコードしたレベルによる分布をまず確認したいはずである。こうした場合、コンディションファイルでは各要因番号をただ括弧に入れて各行一要因の割合で並べるだけでよい（「コード修正なし（no recode)」というケースである)。東京語ゼロマーク化のデータであれば、(6) のようになる。この場合、表 1 で見た要因すべてをそのまま分析に使うことになる。

(6) コード修正なしのコンディションファイル

[13] トークンファイルとコンディションファイルのこの形式は、もともと VARBRUL プログラムの一部が LISP 言語で書かれていたことに依る。

```
(
; Identity recode:  All groups included as is.
(1)
(2)
(3)
(4)
)
```

しかし分析を進めていくと、あるレベルのケースが極端に少なく十分な分析が見込めない場合もあれば、どういうわけかある変数について 1 レベルのケースしかデータ中に見つからなかったケース（VARBRUL 関連資料ではシングルトンファクター (singleton factor) と呼ぶ）、また異なる変数間に従属性が発見された場合などが出てくる。こうした場合、そのままロジスティック回帰に進むわけにはいかないので、適当に変数の再編成をする必要が出てくる。これが「コード修正、リコード (recode)」と呼ばれる作業である。現在入手可能な Goldvarb（Goldvarb 2.1, Goldvarb 2001, Goldvarb X）ではリコードは GUI インターフェースでできるようになっているが（Tokens メニューから Recode Setup）、リコード指定は AND,OR, NOT の 3 関数と、COL 指定および ELSEWHERE, NIL, '\' といった指定語によって可能であるから、手作業でも容易である。例えば名詞句形式差で、WH 代名詞とそれ以外の代名詞を、「代名詞」という括りでまとめたいとする。この場合、コンディションファイルは (7) のようにすればよい:

(7) WH 代名詞とそれ以外の代名詞の合併

```
(
; Merge WH-pro and other pronouns.
(1)
(2 (8 (OR (COL 2 3)(COL 2 8))))
(3)
(4)
)
```

4 行目の指定は、トークンファイルの 2 コラム目（すなわち名詞句形式差要因）が '3'（WH 代名詞）であるか 2 コラム目が '8'（代名詞）であれば（OR）それらのケースをすべて '8' とせよ、ということである。ここで書かれていない場合についてはそのままであるから、実は同じことは以下のように簡潔に書いても良い:

(8) WH 代名詞とそれ以外の代名詞の合併 (2)

```
(
; Merge WH-pro and other pronouns.
(1)
(2 (8 (COL 2 3)))
(3)
(4)
)
```

3つの関数以外は少々説明が必要である。まずCOLは、トークンファイル中のあるコラム（つまりある変数）がある値を撮る場合、という意味である。COL(2 4)であれば、2コラム目が'4'である場合、という意味になる。ELSEWHEREと言うのは、「他の場合はすべて」という意味であり、たとえば2コラム目の変数に4つのレベルa, b, c, dがある場合に、b以下3つを同じコードbにまとめようとする場合、一つ一つ指定したりORでつなぐということをしないで、(9)に見るように簡単に済ませられるという利点がある。

(9) ELSEWHEREの使用例

```
(2 (a (COL 2 a))
   (b (ELSEWHERE)))
```

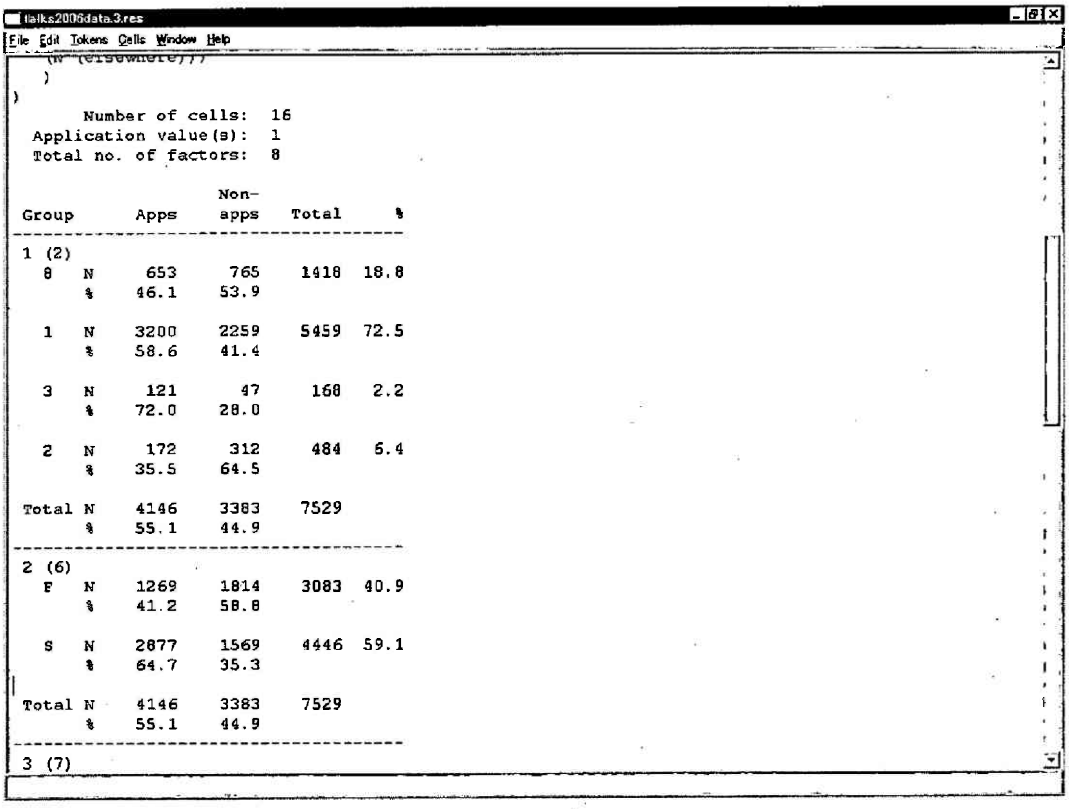
NILというのは、ある変数がある値（レベル）であった場合、それをデータファイルから除外せよ、という意味である。(3 (NIL (COL 3 c)))とすれば、第3コラムの変数が'c'であった場合、以後の分析対象から外すということになる。結局このコマンドでデータファイルを書き換えることになるが、この機能を利用して、トークンファイルを新しく作ることも可能である（Cellsメニューから Recode to New Token file）。これに対して'\'というのは、データ自体は変更しないが、現在おこなっている回帰分析からは無視する、という意味であり、この2つはきちんと区別をする必要がある。

4.3 セルファイル

トークンファイルとコンディションファイルが完成すると、セルファイルが作成できる。セルファイルを作るには、Cellsメニューから Load Cells to Memory を選択し、「現在オープンしているトークンファイルとコンディションファイルからセルを作って良いか？」という確認画面にOKを出し、次の画面でApplication Valueを選択すればよい。Application Valueというのは、従属変数の2つの値のうち、どちらの値の割合を予測したいか、ということだと言い直すことができる。たとえばゼロマーク化の場合、ゼロマーク化が起きた場合を'1'としているが、これは各セルでゼロマーク化が起きた割合に基づいてそれに関わる各要因の重みを計算しようとするのであり、ゼロマーク化が起きない場合（つまり「を」が使われる場合）を基本に考えるのではない、ということである。[14]

[14] ここで *Application Value* と呼ばれているのは、VARBRULが登場した当時の枠組みでは、変異規則（Variable Rule）が適用されて'1'のケースが実現された、と考えられていたからで、"AApplication Value"という呼び名は当時の残滓ということができる。

Application Value を打ち込むと、即座に 2 つのファイルが画面上に現れる。一つは 結果出力ファイルである。これはセルファイル作成以降、クロス表作成 (crosstab)、変数選択付き回帰分析 (stepup and stepdown)、および変数選択なし回帰分析（one-level analysis）といった分析の出力を記録するテキストフォーマットのファイルであり、.res という拡張子を持つ。この段階で、コンディションファイルの内容と各独立変数ごとの度数分布表が打ち出されている（図 3)。



```
talks2006data.3.res
File Edit Tokens Cells Window Help
   )
)
      Number of cells:  16
  Application value(s):  1
  Total no. of factors:  8

                    Non-
Group       Apps    apps   Total      %
-----------------------------------------
1 (2)
  8   N      653     765    1418  18.8
      %     46.1    53.9
  1   N     3200    2259    5459  72.5
      %     58.6    41.4
  3   N      121      47     168   2.2
      %     72.0    28.0
  2   N      172     312     484   6.4
      %     35.5    64.5
Total N     4146    3383    7529
      %     55.1    44.9
-----------------------------------------
2 (6)
  F   N     1269    1814    3083  40.9
      %     41.2    58.8
  S   N     2877    1569    4446  59.1
      %     64.7    35.3
Total N     4146    3383    7529
      %     55.1    44.9
-----------------------------------------
3 (7)
```

図 3: 結果出力ファイル（度数分布表の一部）

変数は Group（factor group という意味である）という見出しで打ち出されているが、1(2) という形式になっている。最初の数字は「最初の独立変数」であることを表し、括弧内の数字は「従属変数を入れればトークンファイルで 2 コラム目に入力されている変数」という意味である。クロス表や回帰分析でもこの最初の数字で変数指定を行うことになるので、この表記法には注意すべきである。また、見出しで Apps となっている所には '1' の場合の件数とパーセンテージが、そして Non-apps の下に '0' の件数とパーセンテージが書き込まれている。

もう一つの出力ファイルがセルファイルであり、ここには従属変数のレベル数、そのコード、独立変数の数とそれぞれの変数の持つラベルの数とそのコードががまず書き出され、その後 '1' と '0' の件数とそのセルの内容（各要因の組み合わせ）がすべてのセルについて記されている（図 4)。つまり、分析に用いるすべての要因を掛け合わせてできる表のセルすべてについてその内容を書き出したものなのであり、よってセルファイル

と呼ばれるわけである。これもファイル形式はテキストフォーマットなので、この形式さえ遵守すれば自分で作成しても構わない。[15]

セルファイル作成中には、(10) の 3 種類のエラーメッセージが出る可能性がある。エラーメッセージが出た場合には、これをクリアしないとクロス表作成、回帰といった次のステップには進めない。

(10) セルファイル作成作業中のエラー

1. コンディションファイルエラー
2. シングルトンファクター (singleton factor)
3. ノックアウトファクター (knockout factor)

コンディションファイルエラーは、文字通りコンディションファイルの指定にミスがあった場合である。この場合セルファイル作成に行く前にエラーメッセージが出る。たいていは括弧の付け忘れが原因であることが多いので、慎重に左右の括弧の数を数えれば防げるミスである。

「シングルトンファクター」についてはすでに触れたが、ある変数に 1 レベルのケースしかデータ中に見つからなかった場合である。この場合、その変数は分析から除く他はないので、この変数の番号を取り除いたコンディションファイルを作成し直して、セルを作り直せばよい。

「ノックアウトファクター」は、変数があるレベルの値を取ると、従属変数がカテゴリカルに '0' か '1' を取る場合である。これはいわばバリエーションがないので、パラメータ値の計算のしようがなく、エラーになってしまう。データ件数が十分多い場合であれば、その値が従属変数の値を大きく左右することを示す証拠であるから、ログにその旨を書いた上で分析からは外すのが良い。データ件数が少数の場合は、やはりデータから除外するか、言語学的に併合が可能な他レベルと合併を考えるのが一般的である。たとえば WH 代名詞とそれ以外の代名詞であれば、「代名詞」として一応は合併する根拠が考えられる。ただし、節と代名詞であれば、2 者のカテゴリーの合併について言語学的に説得力のある議論を展開するのは難しい。ここらについては、ケースバイケースで考える他はないのであり、機械的にまとめるのは絶対に避けるべきである。なお、こうして合併をしてコンディションファイルを変更する場合には、ファイルにセミコロンでコメントアウトして経緯を書いておくと、後々便利である。

4.4 クロス表分析

セルファイルが完成すると、クロス表作成と回帰の 2 つの操作が可能である。最初に Goldvarb を走らせるのであれば、まずはクロス表分析をするのが常道である。クロス表は、`Cells` メニューから `Crosstabulation` を選択すればよい。クロス表を構成する変数番号 2 つと画像 (Picture) とテキスト (Text) というフォーマット選択のウィンドウが出

[15] セルファイルの最後の行は '-1' だけで終わる約束になっている。

図 4: セルファイルの内容

```
11
 3
   48132
   2FS
   2AN
 840 919
1FA
 132 282
1FN
1808 765
1SA
 420 293
1SN
  45 123
2FA
  13  68
2FN
  82  71
2SA
  32  50
2SN
  30  21
3FA
   8   7
3FN
  30  21
3FA
   8   7
3FN
  63  12
3SA
  20   7
3SN
 160 257
8FA
  41 137
8FN
 357 230
8SA
  95 141
8SN
  -1
· 2006/02/18 15:38:55
· Token file: all4.tok
· Conditions: talks2006data.3.cnd
```

たら、適宜指定する。変数番号は、先に .res ファイルで与えられた番号であり、変数番号は列×行の順で半角で入力する。テキストフォーマットの場合そのまま結果出力ファイルに書き込まれ、画像フォーマットは、別ウィンドウで結果が示され、PNG 形式で保存が可能である。名詞句形式差とスタイルのクロス表を画像フォーマットでクロスした結果は、図 5 のようになる。[16]

"talks2006data.3.cel"

- 2006/02/18 15:38:55
- Token file: all4.tok
- Conditions: talks2006data.3.cnd

Group #1 — horizontally.
Group #2 — vertically.

	8	%	1	%	3	%	2	%	Σ	%
F 1	201	34	972	45	38	58	58	23	1269	41
−	394	66	1201	55	28	42	191	77	1814	59
Σ	595		2173		66		249		3083	
S 1	452	55	2228	68	83	81	114	49	2877	65
−	371	45	1058	32	19	19	121	51	1569	35
Σ	823		3286		102		235		4446	
Σ1	653	46	3200	59	121	72	172	36	4146	55
−	765	54	2259	41	47	28	312	64	3383	45
Σ	1418		5459		168		484		7529	

図 5: 名詞句形式差とスタイル差のクロス表出力（画像 (picture) フォーマット）

VARBRUL 分析においてクロス表を作成する目的は、先にコンディションファイルのところで触れた変数間の従属性や、データ分布の極端な偏り、交互作用の存在などをチェックすることにある。SPSS や SAS のような統計パッケージでは、階層モデル (hierarchical model) で分析を行うか特定交互作用項を含んだモデルを指定するかして、交互作用の有意性を検定するという手続きを取る。しかし過去から現在に至る VARBRUL プログラムファミリーでは、いずれも変数間の独立性をいわばデフォルトとして想定している。そこではなはだ原始的な方法だが、交互作用のチェックはまずクロス表で行い、その有意性検定はコンションファイルで交互要因同士を組み合わせた新要因を作成して行うというのが、VARBRUL 分析での手続きなのである。こうした事情もあり、ここで出力できるクロス表は独立変数同士のクロスに限定されている。各々の独立変数と従属変数のクロスは、すでにセルファイルを作成した時に結果出力ファイルに従属変数のレベル別（Apps と Non-apps）度数分布表として書き出されているのである。

また、これも一般の統計パッケージとは異なる点だが、作成できるクロス表は単純ク

[16] この場合の変数指定は 1×2 であることに注意。

ロス表に限定されている。多重クロス表を作成するには、コンディションファイルのところでデータを層別に別ファイルにし(Recode to New Token File)、それぞれ別々にクロス表分析をするという、手間のかかる作業を行うことになる。多重クロスをするのであれば、データをExcelに移してピボット・テーブル分析をした方が良いだろう。

クロス表分析の段階で、変数間の従属性や交互作用の存在が明らかになった場合、再びコンディションファイルに戻り、こうした問題を解決できるように変数を組み直す必要が出てくる。その上で再びセルファイルを作成し、クロス表を作り直して問題がないことを確認した上で、ロジスティック回帰へと進むわけである。ソフトの仕様からは、クロス表を飛び越えてセルファイル作成から一気に回帰へと向かうことも可能だが、上述したような問題を抱えたまま回帰を行っても全く意味がないので、まずは可能な限りクロス表を作成して、問題がないことを確認することが肝要である。

4.5 変数選択付きロジスティック回帰分析

さて、ここまでが無事に終わるといよいよ回帰分析に入ることができる。VARBRULプログラムでは回帰分析に、有意な変数を自動的に選択する変数選択機能付きの分析と、コンディションファイルで指定されたすべての変数をモデルに組み込む変数選択機能なしの、2種類の回帰分析を用意している。

変数選択機能付き分析は、Cellsメニューから Binomial, Up and Down を選択して行う。この段階で、メニュー下方にある Centre Factors オプションを選択すると、各要因のパラメータ値の平均が0.5に設定される。それ以外の場合では、各レベルのパラメータ値は、そのレベルの生起数に応じて重み付けされる。[17] メニューの選択と同時に計算が始まり、結果は結果出力ファイルに書き込まれる。

変数選択は、まず定数項のみから出発して、一つ一つモデルに変数を加えて行き、最終的に有意な変数がなくなるまで続ける変数増加法 (stepup) の段階、そしてすべてを含んだモデルから一つ一つ有意でない変数を削ってゆく変数減少法 (stepdown) の段階まで一気に行われる。各モデルの有意性の判定には対数尤度比が用いられる。具体的には、現在検討中の変数を含んだモデルの対数尤度と、その変数を含まないモデルの対数尤度の差の2倍が、2つのモデルの自由度の差を自由度とするカイ自乗分布に漸近的に従うことを利用して判定される。モデルの自由度は、次のようにして求められる。

(11) モデルの自由度 = (そのモデルに含まれた全レベル (factor) の数)
−(そのモデルに含まれた全変数 (factor group) の数)

それぞれのモデルでこの値を計算し、その差を自由度として2モデルの対数尤度比（差）をカイ自乗値として検定し、増加法であれば有意な場合、有意度の高い変数を取り入れて新しいモデルとし、再び残った変数を一つ一つ検討する作業を、有意な変数がなくなるまで繰り返すわけである。減少法の場合はその逆で、有意でない変数からモデルから落

[17] VARBRULの以前のバージョンでは、この方式でパラメータ値が計算されているので、過去にVARBRULを使って計算された結果と比較する場合などは、このオプションを選択することが望ましい。

としていき、落とす変数がなくなるまでこの過程を繰り返すことになる (Young & Bayley, 1996, 279)。

出力はモデルが含む変数の数によってレベル (level) に点線で区切られ、さらにその中でラン (run) と呼ばれる各段階ごとにまとめられたフォーマットで打ち出される。最初のレベルは定数項 (input probability)[18] すらも検討中であるからレベル 0 であり、以後変数の増減に伴いレベルの数も増減する。ランには通し番号が打たれ、各ランには、モデルの持つセル数、定数項を含んでそのモデルが含む全変数のパラメータ値、対数尤度、そして検討中のモデルと検討中の変数を持たないモデルとの検定結果などが示される。そして各レベルの最後には、そのレベルで選択された変数の番号とレベルが打ち出される。

ゼロマークデータで変数選択をした場合の、増加ステップを見てみよう（図 6–7)。レベル 0 で定数項の計算が終わり、次にレベル 1 の 3 つのランでもっとも対数尤度値が高かったスタイル (FS) が変数として選択され、以後名詞句形式差、隣接性の順に選択されてていることがわかる。今回のデータでは、結局 3 要因すべてが増加法・減少法のいずれでも選択されることになった。

ロジスティック回帰に限らず変数選択についての注意として一般的によく言われるのは、プログラムによる選択結果を鵜呑みにするな、ということである。変数選択機能付き回帰の出力は変数の数が多いと膨大なものとなることもあるが、選択の過程を丹念に追うべきである。新たな変数が加わると共にすでに含まれている変数のパラメータ値が大きく変動するような場合であれば、多重共線性（multicollinearity）や交互作用の可能性を検討すべきであるし、計算が収束 (converge) しない場合があれば、データをもう一度洗い直すべきだろう。

4.6 変数選択なし（ワンレベル）ロジスティック回帰分析

変数選択なしの回帰分析をする目的は、大きく分けて 2 通りである。一つは、何らかの都合である特定モデルの当てはまり具合を見たい場合であり、もう一つは同一変数内のレベルを合併してレベルの有意性を検討したい場合である。後者の場合、合併前モデルと合併後モデルの対数尤度比（差）の 2 倍を計算して変数選択の場合と同様の検定をすることになるが、この合併後のモデルの対数尤度を見るために、この機能を用いるわけである。ロジスティック回帰の目標は、言語学的に納得しうる限りにおいてできるだけ少ない要因で、データへの妥当なあてはまりを示すモデルにたどり着くことである。よって、できるだけコンパクトなモデルに到達できるように、変数全体のみならず、レベル間の有意性も問題になるのである。変数選択機能付きの回帰が、複数ステップにわたって様々なモデルの当てはまり具合を試したのに対して、この回帰プログラムでは単一レベル（しかも単一ラン）に止まるので、「ワンレベル分析 (one-level analysis)」とも呼ばれることがある。

変数選択なしの回帰分析をするには、`Cells` メニューから `Binomial, One Level` を

[18] *input probability* という名も、VARBRUL 構想当時の理論的枠組みに由来する VARBRUL 関係資料独特のものであり、統計学では一般に定数項 (constant) と呼ばれる。変異理論関係の論文では、*grand mean* という言い方も見かける。

図 6: ゼロマーク化データにおける変数選択

```
---------- Level # 0 ----------

Run # 1, 1 cells:
Convergence at Iteration 2
Input 0.551
Log likelihood = -5179.977


---------- Level # 1 ----------

Run # 2, 4 cells:
Convergence at Iteration 5
Input 0.534
Group # 1 -- 8: 0.427, 1: 0.553, 3: 0.692, 2: 0.325
Log likelihood = -5095.358  Significance = 0.000

Run # 3, 2 cells:
Convergence at Iteration 4
Input 0.531
Group # 2 -- F: 0.382, S: 0.618
Log likelihood = -4975.001  Significance = 0.000

Run # 4, 2 cells:
Convergence at Iteration 4
Input 0.511
Group # 3 -- A: 0.574, N: 0.426
Log likelihood = -5119.657  Significance = 0.000

Add Group # 2 with factors FS
```

図 7: ゼロマーク化データにおける変数選択（続き）

```
---------- Level # 2 ----------

Run # 5, 8 cells:
Convergence at Iteration 5
Input 0.518
Group # 1 -- 8: 0.421, 1: 0.548, 3: 0.693, 2: 0.334
Group # 2 -- F: 0.383, S: 0.617
Log likelihood = -4898.722  Significance = 0.000

Run # 6, 4 cells:
Convergence at Iteration 5
Input 0.487
Group # 2 -- F: 0.379, S: 0.621
Group # 3 -- A: 0.581, N: 0.419
Log likelihood = -4906.991  Significance = 0.000

Add Group # 1 with factors 8132

---------- Level # 3 ----------

Run # 7, 16 cells:
Convergence at Iteration 5
Input 0.483
Group # 1 -- 8: 0.423, 1: 0.539, 3: 0.693, 2: 0.341
Group # 2 -- F: 0.380, S: 0.620
Group # 3 -- A: 0.576, N: 0.424
Log likelihood = -4841.723  Significance = 0.000

Add Group # 3 with factors AN

Best stepping up run:  #7
-------------------------------------------
```

選択すればよい。この場合も変数選択付きの場合と同様、Centre Factors オプションが選択可能であり、結果は結果出力ファイルに書き込まれる。変数選択なし回帰分析の出力内容は、各要因のパラメータ値、パーセンテージ、セルごとの実測値（総件数+ゼロマーク化数）・予測値・カイ自乗値、セルごとのカイ自乗値の合計およびそれを層セル数で割った平均値、そして現モデルの対数尤度値である。これら結果出力ファイルに書き込まれる内容とは別に、新たにグラフィックスウィンドウが開いて、セルごとにゼロマーク化の割合の実測値を x 軸に、予測値を y 軸にとった散布図が出てくる（図 8）。各セルはセルの規模に応じた四角形（■）で表示されるが、理想的モデルでは対角線上にすべてのセルが並ぶことになるわけで、このグラフから目視で現在のモデルの当てはまり具合が確認できる仕組みになっている。図 8 のグラフでは、一直線上にほとんどのセルが並んでおり、極めて良いフィットをしていることが分かる。

モデルのデータへの当てはまり具合（フィット (fit) という）は、到達したモデルの吟味には欠かせない資料になる。変異理論では VARBRUL2S の登場以来、セル平均カイ自乗値を重視し、これが 1 以下である場合を良いフィットの目安としている。図 8 の出力でも分かる通り、ゼロマークデータに対して 3 変数を組み込んだモデルでは、この値が 0.4251 とやはりかなり良いフィットを示しており、ここからも、変数選択付き回帰で得られたモデルが、実際にデータにきちんと当てはまる予測を成し遂げていることが確認できる。[19]

では、同一変数間のレベルを合併することの妥当性を実際に検定してみよう。(8) のコンディションファイルを使い、名詞句形式差について WH 代名詞と代名詞を「代名詞」として一つのカテゴリーにまとめて良いかどうかを検討してみよう。これは、ゼロマーク化との関わりに関しては 2 つのカテゴリーに有意差はないという帰無仮説を検討することになる。セルファイルを新たに作成し、ワンレベル分析行うと、対数尤度は-4862.074 であった。合併前モデルが-4838.311 であるので、尤度比カイ自乗値は[20]

$$(12)\quad G^2 = -2 \times (-4862.074 - (-4838.311)) = 47.526$$

のように計算できる。2 モデルはレベル一つの違いなので自由度は 1 であるから、上の結果は $p < 0.001$ と高度に有意な差であることがわかる。よって WH 代名詞とそれ以外の代名詞を「代名詞」としてまとめることは、少なくとも統計的には誤った判断だと言えるわけである。

先にも述べた通り、こうした結果に到達するには実はかなりな分析と時間を要し、そのほとんどはコンディションファイルからクロス表・回帰分析の間のステップに費やされるのが実情である。またその過程では、VARBRUL プログラム以外にも Excel やグラフ

[19] しかしこの統計量にも問題は多く、セル平均カイ自乗値というのも、かなり大まかな目安に過ぎない。ロジスティック回帰でモデルの適合度を見るには、SPSS でもオプションが提供されている従属変数の分類精度、Hosmer and Lemeshow 検定を始め、MacFadden's R^2 などさまざまな統計量が考案されているので、VARBRUL の出力のみに頼らず、こうした統計量を独自に計算するなどしてフィットを検討するべきだろう。それにしてもさまざまな OS 向けに開発がなされており、またロジスティック回帰による分析がデフォルトになった観のある変異理論において、未だにこうした統計量があまり顧みられていないと言うのは、非常に不思議な事態という他はなく、一刻も早く改善されるべきだろう。

[20] G^2 は、尤度比カイ自乗値検定統計量を表す。

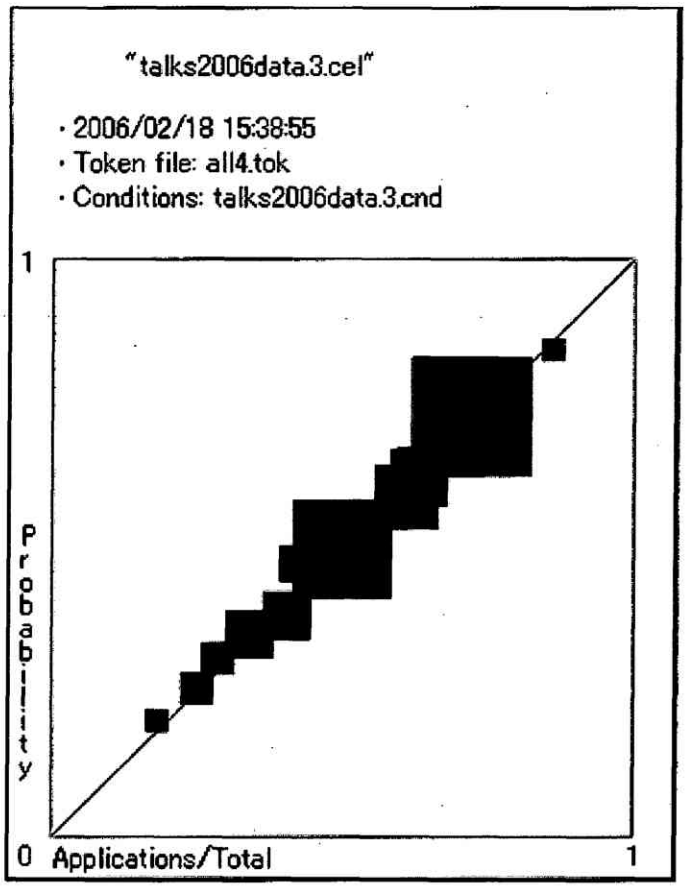


図 8: 変数選択なし回帰分析結果: 散布図

作成ソフトなどを使う必要も出てくる。データ分析に王道はなく、統計的知識のみならず言語学的知識を総動員して、言語学的に意味のある最適モデルの構築に当たらねばならない。VARBRUL はその過程をサポートしてくれる、非常に有用なツールなのである。

5. VARBRUL プログラムの問題点

終わりに、現在リリースされている VARBRUL プログラムで、改良が望まれる点について述べてみよう。まず第 1 の点は、交互作用の取り扱いである。VARBRUL プログラムでは交互作用は、コンディションファイルの操作で関与する要因を組み合わせて新たな要因を作成し、それをモデルに組み入れることで交互作用に対応するのが基本である。この場合交互作用の有意性は、変数選択のステップで検定して判明するわけである。これ以外にも、交互作用を持つ要因のいずれかでデータを分割し、別々の分析をすることで交互作用を回避する方法もあるが (Paolillo, 2002)、いずれにしても交互作用を簡単に指定し、その効果の大きさや有意性の検定まで簡単にしてくれる統計各種パッケージと比較すると、扱いの簡便性に欠ける点は否めない。

第 2 点は、統計オプションの貧困である。VARBRUL には使い勝手を重視した所があり、確かにインターフェースを含めて使い勝手は良いが、その分統計ソフトとしては出

力可能な統計値に今一歩と思われる点が少なくない。たとえば、セルファイル作成の段階で結果出力ファイルに書き込まれる従属変数と各独立変数の表にカイ自乗値を始めとする統計値を出すオプションがあっても良いだろうし、パラメータ値の信頼区間やAICのような情報量基準の組み入れもを出力することが考えられても良いのではないだろうか。フィット統計値については、注19でも触れた通りである。最新のGoldvarb Xとほぼ同内容のVARBRUL2Sがリリースされた1979年当時と現在では、コンピュータの計算能力は桁違なものになっているのだが、出力内容があまり変わらないことに、筆者は強い違和感を感じている。

第3点は、連続変数の扱いである。変異理論では必然的に年齢という変数を扱うことが多いが、VARBRUL2S, Goldvarb 2.1, Goldvarb2001, GoldvarbXといった一般に広く流布したバージョンでは、連続変数をそのまま組み入れることができない。[21] この場合、適当に年齢をいくつかのカテゴリーに区切るしかないのである。年齢を連続変数としてモデルに組み入れたければ、SPSSやSASなどのパッケージを用いる他はない。コンディションファイルにも関わる問題となるだろうが、影響力が大きいソフトなだけに、カテゴリー変数のみというVARBRULの制約は早急に除かれるべきだろう。

参考文献

Cedergren, Henrietta (1973). *Interplay of Social and Linguistic Factors in Panama*. Ph.D. thesis, Cornell University.

Cedergren, Henrietta & Sankoff, David (1974). Variable Rules: Performance as a Statistical Reflection of Competence. *Language*, **50**, 333–355.

エヴェリットB.S. (1980).『質的データの解析 — カイ二乗検定とその展開』. 新曜社. 山内光哉 監訳, 弓野憲一・菱谷晋介訳.

Everitt, B.S. (1992). *The Analysis of Contingency Tables* (Second edition). Chapman & Hall/CRC.

Fienberg, Stephen E. (1980). *The Analysis of Cross-Classified Categorical Data* (Second edition). MIT Press.

Haberman, S.J. (1973). The Analysis of Residuals in Cross-Classified Tables. *Biometrics*, **29**, 205–220.

浜島信之 (2000).『多変量解析による臨床研究』(3版). 名古屋大学出版会.

石村貞夫 (2001).『SPSSによる多変量データ解析の手順』. 東京図書.

[21] VARBRUL3では連続変数の扱いも可能であり、また従属変数が3値の場合（multinomial）も扱えることになっている (Rousseau & Sankoff, 1978)。この機能を使った研究例としては、片岡 (2004), Kataoka (2005) を参照。

片岡邦好 (2004). 空間指示枠の変異について: VARBRUL 分析による一考察. 『社会言語科学会第 14 回大会論文集』, pp. 229–230.

Kataoka, Kuniyoshi (2005). Variability of spatial frames of reference in wayfinding discourse on commercial signboards. *Language in Society*, **34** (4), 593–632.

Labov, William (1994). *Principles of Linguistic Change*, Vol. 1. Blackwell Publishers.

前川喜久雄 (2001). 音声分野における統計的研究手法—音声の変異をめぐって—. 『日本語学』, **20**, 144–156. 4 月臨時増刊号.

増山幹高・山田真裕 (2004). 『計量政治分析入門』. 東京大学出版会.

松田謙次郎 (2000). 東京方言格助詞「を」の使用に関わる言語的諸要因の数量的検証. 『国語学』, **51** (1), 61–76.

Matsuda, Kenjiro (1995). *Variable Zero-marking of (o) in Tokyo Japanese*. Ph.D. thesis, University of Pennsylvania.

西平重喜 (1985). 『統計調査法』 (2 版). 培風館.

Paolillo, John C. (2002). *Analyzing Linguistic Variation: Statistical Models and Methods*. CSLI Publications.

Rietveld, Toni & van Hout, Roeland (1993). *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

Rousseau, Pascale & Sankoff, David (1978). Advances in Variable Rule Methodology. In Sankoff, David (Ed.), *Linguistic Variation: Models and Methods*, pp. 57–69. Academic Press.

Simpson, E.H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of Royal Statistical Society Series B*, **13**, 238–241.

高橋善弥太 (1995). 『医者のためのロジスチック・Cox 回帰入門』. 日本医学館.

丹後俊郎, 山岡和枝, 高木晴良 (1996). 『ロジスティック回帰分析—SAS を利用した統計解析の実際—』. 朝倉書店.

内田治 (2000). 『よくわかる EXCEL による統計解析』 (2 版). 東京図書.

Young, Richard & Bayley, Robert (1996). VARBRUL Analysis for Second Language Acquistion Research. In Bayley, Robert & Preston, Dennis R. (Eds.), *Second Language Acquisition and Linguistic Variation*, pp. 253–306. John Benjamins.

Author's E-mail Address: kenjiro@sils.shoin.ac.jp
Author's web site: http://sils.shoin.ac.jp/~kenjiro/